

The Residual Discontinuity Test for p -Hacking

Tomas Havranek

Charles University • Meta-Research Innovation Center at Stanford



Zuzana Irsova

Charles University



Pedro Bom

University of Deusto



Heiko Rachinger

Univ. of the Balearic Islands

MAER-Net Colloquium 2026 • Chemnitz, Germany

Reported precision is a choice

Table: Regression of test scores on class size

	Krueger (1999)	Replications using different computations of SE				
	(1)	(2) Bootstrap of class clusters	(3) Class clusters	(4) School clusters	(5) Huber-White SE	(6) Plain vanilla SE
Small class	4.82 (2.19)	4.82 (2.00)	4.82 (1.88)	4.82 (1.38)	4.82 (0.79)	4.82 (0.76)
Sample	5,861	5,861	5,861	5,861	5,861	5,861

Notes: Dependent variable: test score percentile. Standard errors (SE) in parentheses.

Meta-analysis of the class size effect (JOLE)

Available at meta-analysis.cz/class. Spoiler: effect size tiny.

Changing controls changes precision

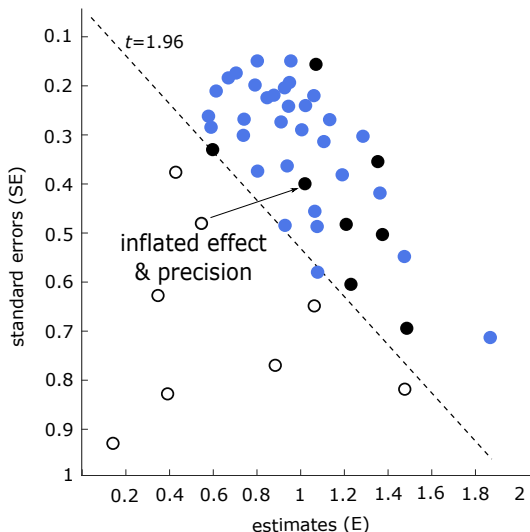
Table: Regression of test scores on class size

	(1)	(2)	(3)	(4)
Small class (treatment)	4.82 (2.19)	5.37 (1.26)	5.36 (1.21)	5.37 (1.19)
White/Asian			8.35 (1.35)	8.44 (1.36)
Girl			4.48 (0.63)	4.39 (0.63)
Free lunch			-13.15 (0.77)	-13.07 (0.77)
White teacher				-0.57 (2.1)
Teacher experience				0.26 (0.10)
Master's degree				-0.51 (1.06)
School intercepts	No	Yes	Yes	Yes
Sample	5,861	5,861	5,861	5,861

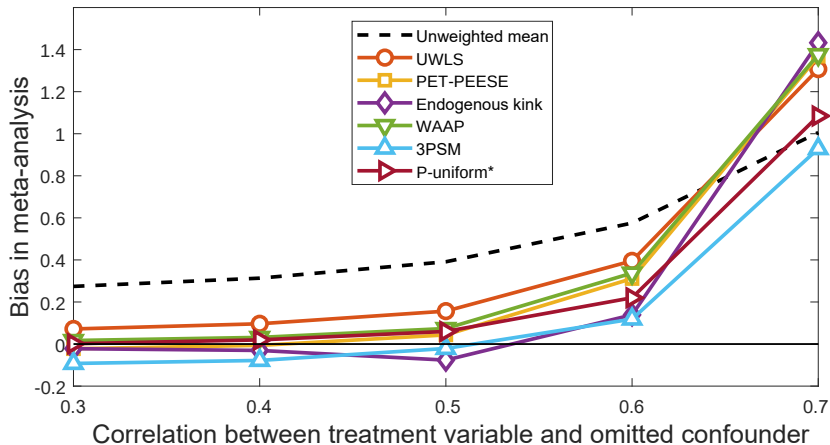
Notes: Adapted from Krueger (1999). Dependent variable: test score percentile.

Standard errors in parentheses.

P-hacking possible on both effect and precision



Cures worse than the disease?



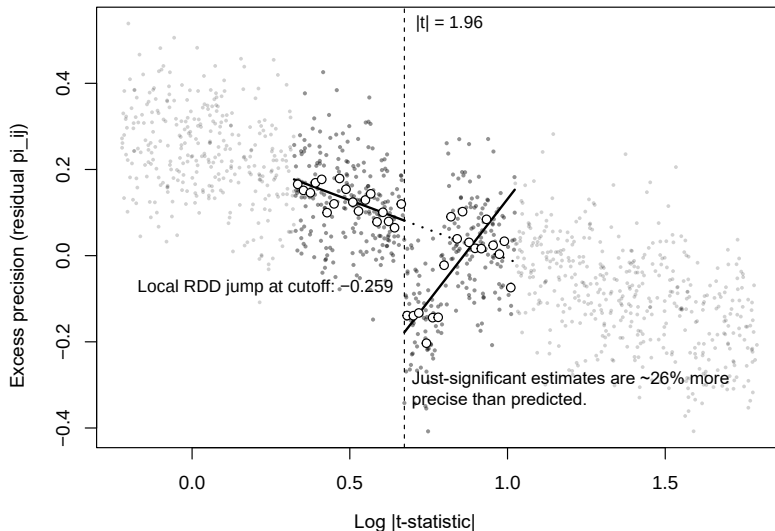
Diagnostic: excess precision

Residual from meta-regression:

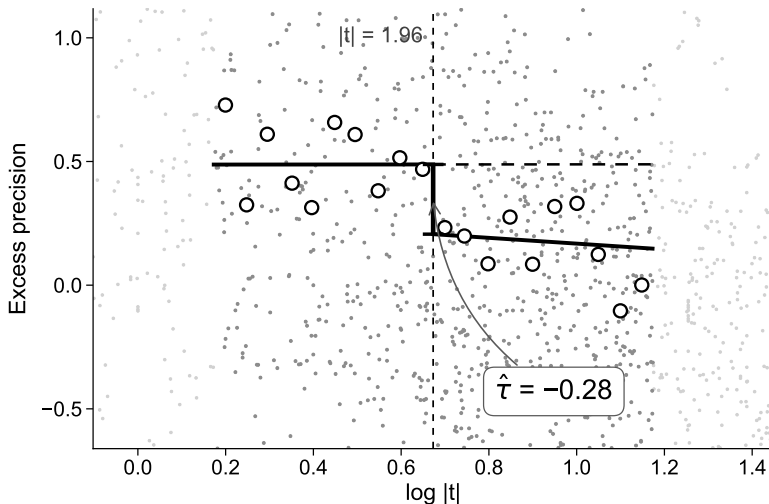
$$\log \text{SE}(\hat{E}_i) = \alpha_0 + \alpha_1 \log N_i + \underbrace{\pi_i}_{<0: \text{excess precision}}$$

- Negative $\pi_i \rightarrow$ reported **SE too small relative to N** $\rightarrow$ may indicate p -hacking.
- Interpretation: $\pi_i = -0.25$ means SE is roughly one quarter smaller than expected.

Residual discontinuity test (simulated)



Pilot: excess sensitivity of consumption



Outside the top-5 journals; 1,039 estimates, 90 studies in the window. Axis trimmed.

Does the picture survive RDD checks?

- Placebo cutoffs 1.5, 2.5: $p = 0.73, 0.79$
- Drop the points on 1.96: $\hat{\tau}$ between -0.30 and -0.23
- Bandwidth $\times 0.5$ to $\times 1.5$: -0.32 to -0.18
- Sample size, publication year: **smooth through the cutoff**

Weighted Adjustment Instrumental Variable Estimator

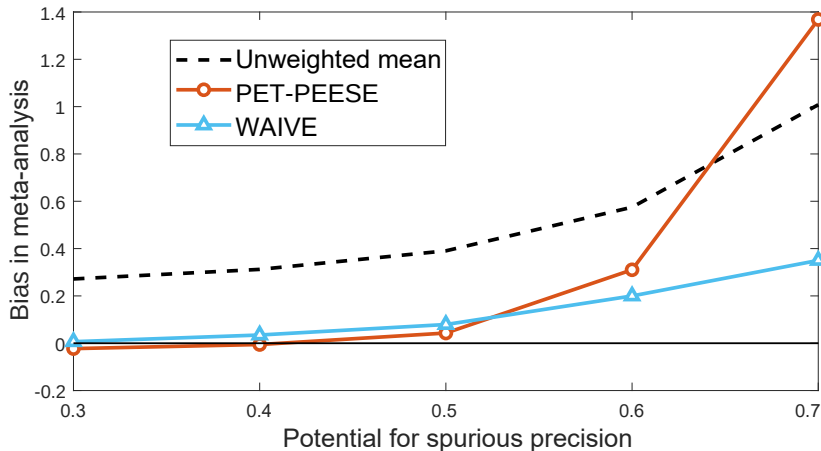
WAIVE combines two adjustments:

- 1 Replace reported SEs by the component predicted from sample size (instrument). This is MAIVE.
- 2 Downweight spuriously precise estimates **near significance cutoffs**. Exponential tilt:

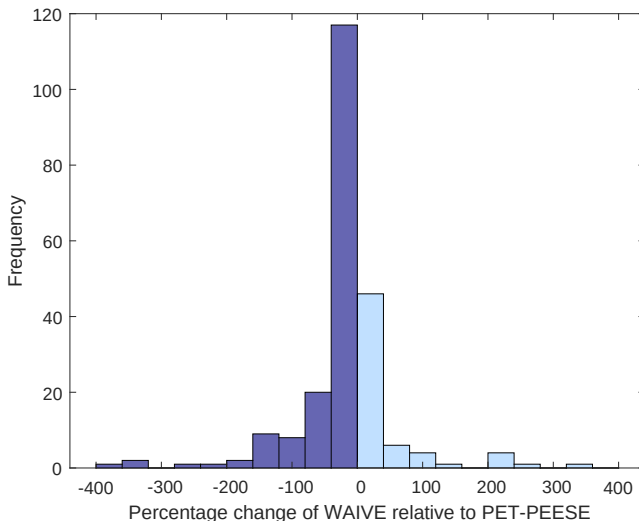
$$\omega_i = \exp[-D_i \max(0, -\pi_i)],$$

- D_i is high when $|t_i|$ is close to 1.96.
- Weight $\omega_i \in (0, 1]$; only excess precision penalized.

WAIVE alleviates the bias in simulations



WAIVE shrinks PET-PEESE in applications



Practical takeaway

The test

RDT: Excess precision should not **jump at $|t| = 1.96$** .

If it does

MAIVE instruments the standard error with sample size.

WAIVE downweights what is left (more aggressive correction).

Try it

EasyMeta.org · meta-analysis.cz/maive

Papers on instrumental meta



Irsova Z., P. Bom, T. Havranek, & H. Rachinger (2025):
Spurious Precision in Meta-Analysis of Observational
Research.

Nature Communications 16: 8454.



Irsova Z., T. Havranek, X. Bortnikova, & F. Bartos (2026):
Meta-Analysis of Field Studies on Beauty and Professional
Success.

Nature Human Behaviour, R&R.



Cala P., T. Havranek, Z. Irsova, M. Luskova, J. Matousek
& J. Novak (2026): Financial Incentives and Performance:
A Meta-Analysis of Experiments in Economics.

Journal of Political Economy Microeconomics, forthcoming.

Backup slides for discussion.

Where to look for more

- **Method paper.** Irsova, Bom, Havranek & Rachinger (2025), *Nature Communications* 16:8454.
- **Tools and replication.** easymeta.org · meta-analysis.cz/maive · MAIVE package on CRAN

Excess precision: log first-stage residual

First stage (log form; extension over published MAIVE)

$$\log(\text{SE}_i) = \alpha + \beta \log(N_i) + \gamma' X_i + \pi_i.$$

- $\pi_i < 0 \Rightarrow$ reported SE **smaller than sample size predicts**.
- X_i : pre-specified design features; *not* researcher-chosen inference tweaks.
- Interpretation: $\pi_i = -0.10 \Rightarrow \text{SE} \approx 10\%$ below prediction (exact: $1 - e^{-0.10} = 9.5\%$).

First stage from MAIVE (Nature Comms 2025); log version used going forward.

Downweighting: D_i on $\log |t|$

Triangular danger-zone kernel (default)

$$D_i = \max\left\{0, 1 - \frac{|\log |t_i| - \log(1.96)|}{h}\right\}, \quad t_i = \hat{E}_i/\text{SE}_i.$$

- **Compact support:** $D_i = 0$ outside bandwidth h ; same scale as RDT.

WAIVE weight

$$\omega_i = \exp[-D_i \max(0, -\pi_i)] \in (0, 1].$$

Penalizes only estimates with excess precision *and* $|t|$ near the cutoff.

RDT: residual discontinuity at the significance cutoff

Test of continuity in $E[\pi_i \mid r_i]$ at $r = c$

$$r_i = \log |t_i|, \quad c = \log(1.96),$$
$$\tau_{\text{RDT}} = \lim_{r \downarrow c} E[\pi_i \mid r_i = r] - \lim_{r \uparrow c} E[\pi_i \mid r_i = r].$$

- Null: $E[\pi_i \mid r_i]$ **continuous** at c . Holds if (i) selection works only through the reported t , which reweights the *density* of r and not the conditional mean; and (ii) $E[\pi_i \mid r_i]$ is smooth absent manipulation.
- $\tau_{\text{RDT}} < 0$: just-significant SEs drop below the left-side extrapolation.
- Same scale as π : $\tau = -0.10$ means SEs about a tenth smaller than continuity implies.

RDT: estimation and falsification battery

Estimation

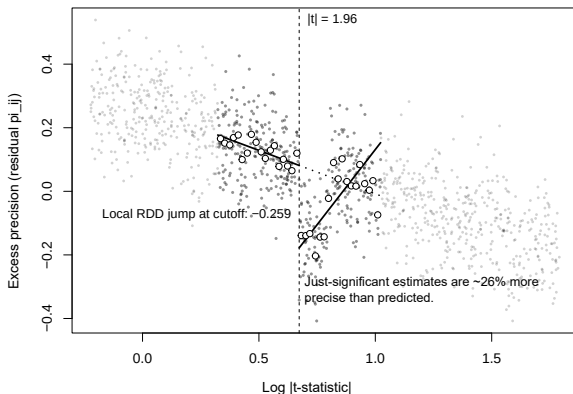
- Outcome: $\hat{\pi}_i$
- Running variable: $\log |t_i|$
- Local linear, both sides
- Triangular kernel
- Bandwidth $h = 0.50$; MSE-optimal as a check
- Cluster bootstrap by study

Falsification battery

- Donut hole around $|t| = 1.96$, **sized to the reported digits of SE**
- Placebo cutoffs ($|t| = 1.5, 2.5$), one-sided windows
- Predetermined covariates continuous at c
- Bandwidth sensitivity
- Density tests (McCrary; Cattaneo et al.), for context

No global polynomial; identification is intentionally local.

RDT: the picture (simulated)



Simulated: a -0.22 step injected just above the cutoff. The fitted jump reads -0.259 because a straight line through a narrow step overshoots it.

Pilot: sensitivity to the estimator

Same subsample, other estimators

- Cluster bootstrap at $h = 0.50$ (the slide): -0.28 , $p = 0.026$.
- `rdrobust`, same h : $+0.05$; MSE-optimal $h = 0.84$: -0.06 .
- At the MSE-optimal h the window spans $|t| \approx 1$ to 4.5 and averages it away.

Rounding of SE: the donut has to match the digits

- SE enters **both** the outcome (+) and the running variable (−), so rounding error δ moves them in opposite directions.
- With ordinary publication selection this lifts $E[\pi \mid r]$ just *below* c more than just above: a **negative jump from rounding alone**, over a width equal to the rounding scale.
- An SE printed to **two decimals** as 0.03 carries $\delta \approx 0.1$ in logs.
- A donut of $|t| \in [1.95, 1.97]$ matches rounding in t . Rounding in SE is wider: **size the donut from the reported digits**, or use a reported t where one exists.

Simulation: data-generating process

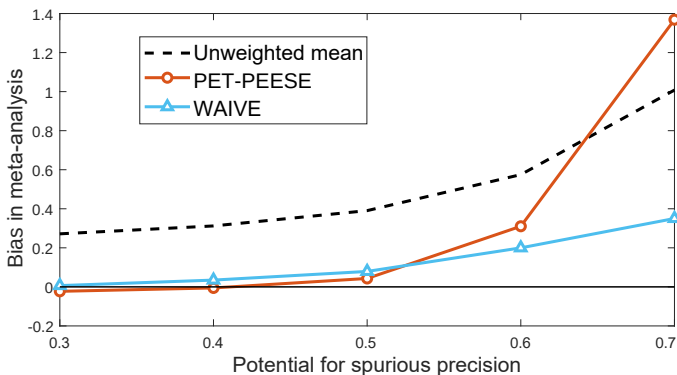
Primary studies

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + u, \quad X_2 = \psi X_1 + (1 - \psi) \varepsilon.$$

- True effect $\alpha_1 = 1$ (bias on the same scale).
- X_2 : true control; replacing it with weaker proxies creates OVB, **stronger as ψ rises**.
- $M = 80$ studies; sample sizes calibrated to published meta-analyses.
- 50% potential p -hackers; search over proxy controls until $t > 1.96$ or $H = 50$ attempts.

Simulation design: MAIVE (Nature Comms 2025); tools at easymeta.org.

Simulation: WAIVE alleviates the bias



At the top of the range WAIVE holds bias to under a third of PET-PEESE.
WAIVE is still experimental; simulation design from MAIVE (2025).